SUPPLEMENTAL INFORMATION

EXTENDED EXPERIMENTAL PROCEDURES

Ethics statement. Subjects were recruited for this study using protocols approved by human subjects committees at Tulane University, Harvard University, Broad Institute, Irrua Specialist Teaching Hospital (ISTH), Kenema Government Hospital (KGH), Oyo State Ministry of Health, Ibadan, Nigeria and Sierra Leone Ministry of Health. All patients were treated with a similar standard of care and were offered the drug Ribavirin, whether or not they decided to participate in the study. Treatment with Ribavirin followed the currently recommended guidelines (McCormick et al., 1986) and was generally offered as soon as Lassa fever (LF) was strongly suspected.

Sample collections, study subjects. Human samples were obtained from patients with LF; all samples were acquired on the day of admission before any treatment regimens had been started. Ten ml of whole blood was collected and plasma or serum was prepared by centrifugation at 2500 rpm for 15 minutes. Diagnostic tests for the presence of LASV were performed on-site using PCR (Olschlager et al., 2010) and/or ELISA antigen capture assays (Branco et al., 2011b). Both assays have comparable sensitivity (Branco et al., 2011a). All samples were re-tested by PCR upon receipt at Harvard University; in general ~60% of samples from both Nigeria and Sierra Leone that tested positive in the field also tested positive in the US. Only samples that tested positive for LASV in the field and by PCR at Harvard University were used in this study. Rodents (all from Sierra Leone) were trapped in case-households, humanely sacrificed and samples were collected from serum and/or spleen.

All samples were inactivated in either Buffer AVL (Qiagen) or TRIzol (Life Technologies) following the manufacturer's standard protocols and stored in solar-driven −20° C freezers. In later samples, carrier RNA was omitted from Buffer AVL as we found that this significantly improved the efficiency of library construction (Matranga et al., 2014; Stremlau et al., 2015). Every three to six months inactivated samples were shipped on dry ice to Harvard University where samples were kept at −80° C until further processing.

Case-fatality rate calculations. Case-fatality rates (CFRs) were calculated from patients from Sierra Leone and Nigeria that had all of the following characteristics: (i) we knew the outcome of their illness, (ii) they tested positive for LASV in the field, (iii) their samples were confirmed positive by PCR upon retesting at Harvard University, and (iv) their samples were successfully sequenced at the Broad Institute. In Sierra Leone the CFRs were calculated from LASV positive patients that came to KGH from 2009 – 2013. In Nigeria they were calculated from LASV positive patients that came to ISTH from 2009 – 2012.

Ebola virus dataset. Consensus sequences and raw data for all the EBOV samples produced by our group from Sierra Leone are available via NCBI BioGroup PRJNA257197 (Gire et al., 2014). Sequences from Guinea (KJ660346, KJ660347, KJ660348), Liberia (KP178538), and Mali (KP260799, KP260800, KP260801, KP260802) were downloaded directly from NCBI. All accession numbers can be found in Table S2.

Viral RNA isolation and QC. AVL-inactivated RNA was isolated using the QIAamp Viral RNA Minikit (Qiagen) according to the manufacturer's protocol, except that 0.1 M final concentration of β-mercaptoethanol was added to each sample to dissolve any agglutinated products (Matranga et al., 2014). TRIzol-inactivated RNA was isolated according to the manufacturer's protocol with slight modifications. Briefly, 200 µl 1-bromo-2 chloropropane (BCP) was added for every 1 ml TRIzol used for inactivation. After phase separation, 20 µg of linear acrylamide was added to the aqueous portion. All extracted RNA was resuspended in water and treated with Turbo DNAse (Ambion) to ensure removal of contaminating DNA.

Quantification of LASV and EBOV genome load. We were unable to calculate traditional titers for LASV and EBOV due to their BL-4 status. Instead, LASV/EBOV RNA and host 18S rRNA were quantified using the Power SYBR Green RNA-to-Ct 1-Step qRT-PCR assay (Life Technologies). For LASV, primers were designed to produce short (~60-90 bp) amplicons within the LASV NP gene for both Sierra Leone-like strains (Table S1, primers tab, SL1, SL2) and Nigerian-like strains (Table S1, primers tab, NG1, NG2). Standard PCR amplicons encompassing qRT-PCR products were prepared to determine viral copy number in qRT-PCR assays. This was done by using synthetic oligonucleotides representing a portion of the LASV S segment within the NP gene as a template for PCR (Table S1, primers tab, SL3-5, NG3-5). These amplicons were cleaned up using Agencourt AMPure XT beads (Beckman Coulter Genomics) and quantified by Quant-iT (Invitrogen). Amplicon concentrations were converted to LASV copies/µl for quantification purposes. Host RNA was quantified using human 18S rRNA primers using genomic DNA (Promega) as a standard control (Table S1, primers tab, 18S1, 18S2). For EBOV, we used data from the Gire *et al.* publication (Gire et al., 2014). All reactions were performed on the ABI 7900HT (Applied Biosystems). To control for differences in degradation, RNA extraction, and sample handling, the amount of LASV/EBOV material was normalized to that of host 18S rRNA material to give a final log ratio of viral copies/µl to 18S rRNA copies/µl.

Selective RNA depletion. Samples with ≥ 1 detectable copy of LASV RNA were used for selective depletion. Briefly, prior to library construction, poly(rA) carrier and host ribosomal RNA was depleted from LASV samples using an RNase H based selective depletion approach (Adiconis et al., 2013). Oligo(dT) and DNA probes complementary to human rRNA were hybridized to the sample RNA. The sample was then treated with RNase H (Hybridase, Epicentre). All synthetic DNA oligos were removed by treating with RNase-free DNase (Qiagen) according to the manufacturer's protocol. Carrier- and rRNA-depleted samples were purified using 1.8x volumes AMPure RNA clean beads (Beckman Coulter Genomics) and eluted into water for cDNA synthesis.

Illumina library construction and sequencing. cDNA synthesis and Illumina paired-end library construction were done similarly to published RNA-Seq methods (Matranga et al., 2014; Levin et al., 2010). Briefly, controls were used to monitor our library construction process. We spiked in 1 pg of one, unique synthetic RNA (ERCC, gift from M. Salit, National Institute of Standards and Technology (Jiang et al., 2011)) using a different RNA for each individual LASV sample to aid in tracking our viral sequencing process and potential index cross-contamination. Also, we prepared libraries from 200 ng human K-562 total RNA (Ambion) with each batch as a LASV negative control. RNA samples were fragmented for 4 minutes at 85° C using NEBNext Fragmentation buffer (New England Biolabs). After fragmentation, samples were purified using 2.2x volume AMPure RNA clean beads (Beckman Coulter Genomics). The fragmentation step was removed for libraries prepared using Nextera XT tagmentation (Illumina, see below). Next, double stranded cDNA was prepared using randomly-primed reverse transcription (first strand) and replacement synthesis (second strand). For adapter-ligated Illumina libraries (~50% of libr used up to 18 cycles of PCR to generate our libraries. For Nextera XT tagmentation libraries (remaining 50% of libraries), we used 16 cycles of PCR to generate our libraries. Finally, we used qPCR to calculate copies of LASV in the final libraries. Each individual sample was indexed with an 8 bp unique barcode (or dually-barcoded in Nextera XT libraries) and libraries were pooled equally and sequenced on the HiSeq2000 (101 bp paired-end reads; Illumina), the HiSeq2500 (101 or 150 bp paired-end reads; Illumina), or the MiSeq (150 bp paired-end reads; Illumina) platforms. All the data were deposited at NCBI under BioProject PRJNA254017.

Library preparation, sequencing, and assembly of EBOV samples have previously been described and performed (Gire et al., 2014; Matranga et al., 2014). Since EBOV samples experienced a much shorter time frame between collection and sequencing (Gire et al., 2014), we found that EBOV sequencing led to a higher percentage of viral reads (Gire et al., 2014; Matranga et al., 2014) than LASV sequencing (Figure S1A). However, the normalized genome abundance was very similar for EVD and LF patients from KGH (Figure 4A).

Demultiplexing of sequencing runs and QC. Raw sequencing reads were demultiplexed using the Picard v1.4 pipeline (Institute, 2012b) and saved as BAM files (Li et al., 2009). To avoid barcode cross-contamination between samples the default settings were changed to allow for no mismatches in the barcode and a minimum quality score of Q15 in the individual bases of the index. Sequencing quality metrics were collected using FastQC v0.10.0 (Institute, 2012a) and only high-quality sequencing libraries were used in subsequent analyses.

Assembly of *M. natalensis* transcriptome. Eleven sequencing runs from eight individual rodent samples were pooled. Low quality reads and bases were removed with Trimmomatic v0.15 (Lohse et al., 2012) using the following parameters: LEADING:20 TRAILING:20 SLIDINGWINDOW:4:25 MINLEN:70. BMTagger v3.101 (Agarwala and Rotmistrovsky) was then used to remove all ‘contaminating’ reads (primarily E. coli, spike-in RNAs and LASV) and duplicates were removed using PRINSEQ-lite v0.19.3 (Schmieder and Edwards, 2011). Paired-end de novo assembly of the cleaned reads was performed using Trinity r2011-11-26 with a minimum contig length of 1,000. Out of the resultant contigs > 99% were classified as being of rodent origins using BLASTn v2.2.24+ (Figure S1D). Next, ORFs were predicted using EMBOSS getorf v6.2.0 (Rice et al., 2000) with a minimum size cutoff of 600. The ORFs were clustered using USEARCH v6.0.203 (Edgar, 2010) with an identity threshold of 0.95 and a representative sequence from each cluster was selected (the 'centroid'). The centroids (n = 7,028) are available in File S1, and were used in all downstream analyses.

Assembly of full-length LASV genomes. BAM files were converted to Fastq format and LASV reads were extracted using Lastal r247 (CBRC, 2012) with a custom-made database containing full-length arenavirus genomes. The filtered reads were *de novo* assembled using Trinity r2011-11-26 with a minimum contig size of 300 (Grabherr et al., 2011). The assembly pipeline can be accessed on GitHub (https://github.com/broadinstitute/viral-ngs). If the sequenced sample contained too few reads for efficient *de novo* assembly (usually when average depth of coverage < 20x) a hybrid between *de novo* assembly with Trinity and a manual alignment-based approach implemented in Geneious was used instead ('Assisted assembly', Table S1). Once contigs had been generated, all sequencing reads from individual samples were aligned back to its own LASV consensus using Novoalign v2.08.02 (Novocraft, 2012) with the following stringent parameters -k -l 40 -g 40 -x 20 -t 100. Duplicates were removed using Picard v1.4 and BAM files were subjected to local realignment using GATK v2.1 (McKenna et al., 2010). If multiple sequencing runs had been performed for the same sample, BAM files were merged using Picard v1.4 before further analyses. Consensus sequences were called from the LASV-aligned reads using samtools v0.1.19 (Li et al., 2009) with the following parameters: samtools mpileup -Q 15 -uB -q 1 -d 10000 -f reference.fasta sample.bam. bcftools view -Acg -. vcfutils.pl vcf2fq > sample-consensus.fastq. All generated genomes were manually inspected, checked and corrected for accuracy, such as the presence of intact ORFs, using Geneious v6.1. Regions were depth of coverage was less < 3x were called as 'N'. Samples that failed to generate high-quality consensus sequences were excluded from all further analyses. We found that consensus sequences for the individual LASV genomes were consistent across technical replicates, sequencing machines (MiSeq, HiSeq2000, and HiSeq2500), library preparation methods, and laboratories used for library generation (Harvard University or the Broad Institute).

Multiple sequence alignments. Consensus sequences were aligned using MAFFT v6.902b (Katoh et al., 2002) with the following parameters (L-INS-i): --localpair --maxiterate 1000 --reorder --ep 0.123 before being trimmed using trimAl v1.4 (Capella-Gutierrez et al., 2009) with the maximum likelihood specific parameter: -automated1. Typically this would remove ~ 1-2% of the positions in the alignments. Codon-based alignments - which were used for the majority of our analyses - were then generated using MAFFT as implemented in Geneious v6.1 (Drummond et al., 2010). Key alignments can be found in File S1.

Recombination and reassortment analyses. Alignments containing either all sequences, Nigerian sequences or Sierra Leonean sequences were screened for recombinant viral strains using the programs RDP, GENECONV, MAXCHI, CHIMAERA, 3SEQ, BOOTSCAN and SISCAN as implemented in the RDP3 software package (Martin et al., 2010). Defaults settings were used, except for the following changes: RDP, window size: 100 bp; BOOTSCAN, number of bootstrap replicates: 100, window size: 100 bp, step size: 20 bp, model options: Felsenstein, 1984; MAXCHI, variable window size was used, strip gap was selected; MAXCHI and CHIMAERA variable sites per window set to 50; SISCAN, window size: 100 bp with step size 20 bp. Potential recombinant sequences were identified when three or more methods were in agreement with *P*-values of < 0.001 (Bonferroni corrected). No recombinant sequences were identified in any of our screens as we found no evidence for phylogenetic incongruence in our datasets.

To detect reassortment between L and S segments two different approaches were used. Firstly, concatenated alignments of the L and S segments from Nigeria or Sierra Leone were scanned for breakpoints between the two segments using RDP3 as described above. A reassorted strain was identified when four or more methods identified the breakpoints with *P*-values of < 0.001 (Bonferroni corrected). Secondly, the reassortments were confirmed using the program GiRaF v1.02 with default settings (Nagarajan and Kingsford, 2011) and only samples for which GiRaF and RDP agreed on the identified reassortant were called.

Phylogenetic tree reconstruction. Maximum likelihood phylogenies were made with RAxML v7.3.0 (Stamatakis et al., 2005) using the GTRγ nucleotide substitution model. Fifty consecutive runs were performed and the tree with the best likelihood score was bootstrapped with 500 pseudoreplicates. All trees were rooted using the 1969 Pinneo strain of LASV (Bowen et al., 2000). To confirm the topology of the trees, and to verify Pinneo as the correct root of the tree, we also created alignments and trees using Mopeia virus (GenBank ID: DQ328874.1) as an outgroup. In all cases did we find that the topology of the trees were in agreement (File S1).

Estimation of human-to-human transmission. Human LF patient samples from Sierra Leone collected in 2012 with detailed collection dates (day, month, year) were used for the transmission analysis of LASV. For EBOV, samples from Guinea, Sierra Leone, Liberia and Mali were used. All the included samples can be found in Table S2. Maximum likelihood phylogenetic trees were created for the LASV S segment (n = 21) and EBOV (n = 131). For LASV, a strain from Liberia (G1200) was initially included and the tree rooted on that strain; this sample was then removed from subsequent analyses, alignments recreated, and the tree was rooted on the nearest neighbor to G1200 (G2259). The EBOV tree was rooted on the three strains from Guinea (KJ660346, KJ660347, and KJ660348). Root-to-tip distances were calculated using the program PATH-O-GEN and plotted against the collection dates. R^2 and *P*-values were calculated based on linear regression.

Molecular dating using BEAST. Phylogenies incorporating time of sampling were estimated using Bayesian Markov Chain Monte Carlo (MCMC) as incorporated into the program BEAST v1.7.4 (Drummond and Rambaut, 2007). Several evolutionary models (HKY, GTR, SRD06), clock models (lognormal relaxed, exponential relaxed, strict), codon partitioning (no partitioning, 1,2,3 partitioning, [1+2],3 partitioning) and population sizes (constant, exponential, Bayesian skyline) (Drummond et al., 2005; Drummond et al., 2006) were tested, all with very similar results (Figure S3A, B and Table S2). Model likelihoods were estimated using an importance sampling estimator (Suchard et al., 2003) and Bayes factor analysis was used to select the best parameters (Li and Drummond, 2012). All these analyses were performed on the Batch 1 dataset (Figure 1C). For the final analyses - using the Matched dataset (Figure 1C) - a model incorporating a lognormal relaxed clock, exponential growth, HKYγi with four categories and codon partitioning (srd06) was run for 100 million generations, sampled every 1000 generations using a 10% burn-in rate until all statistics had ESS values > 1,000 (File S1). Maximum-clade credibility trees summarizing all MCMC samples were generated using TreeAnnotator v1.7.4

with a burn-in rate of 10%. To examine the effect of positive selection on the molecular dating estimates, branch lengths were also estimated using the branch site random effects likelihood model in HyPhy, with three dN/dS classes fit for each branch (Kosakovsky Pond et al., 2011; Pond et al., 2005).

Codon adaptation analysis. The synonymous codon usage and sequence composition of each of the full-length LASV consensus sequences in the Matched dataset, or the EBOV sequences from the Gire *et al.* dataset (Gire et al., 2014), were compared to that of human, *M. natalensis*, mouse, and chimpanzee. Each ORF identified in the *M. natalensis* transcriptome was queried against the human genome using BLASTn v2.2.24+ with default settings, and the best hit was then matched to orthologs in the other genomes using ENSEMBL (Birney et al., 2004). In cases where multiple isoforms existed for an ENSEMBL ortholog in a given genome, only the longest isoform was retained. The codon sequences of genes with a 'complete set' of orthologs, represented in all four mammals, were then aligned with PRANK v10.08.02 with the -translate option and default settings (Loytynoja and Goldman, 2005) and all gaps were trimmed, such that every orthologous codon was represented in each mammalian sequence. Codon frequencies and codon usage tables for each mammal were calculated from the resulting alignment of 2,489 orthologs (692,522 codons). The codon adaptation index (CAI) of each full-length viral genome (coding sequences only) was computed separately relative to each host codon usage table using CAIcal v1.4 (Puigbo et al., 2008a). CAI was normalized by the 'expected neutral CAI,' based on 500 randomized viral sequences, conserving the observed GC content and amino acid usage (Puigbo et al., 2008b). The results are interpreted such that a value above ‘1’ is higher than neutral and for all LASV strains analyzed these values were statistically significant.

To account for potential biases in selecting an ortholog dataset based on M. natalensis samples sequenced from Sierra Leone, we also calculated CAI of the individual LASV sequences using full gene-sets from humans (total ORFs ~19,000) and M. natalensis (assembled ORFs = 7,028). These results were in full agreement, and in both cases we observed that Sierra Leone LASV strains had significantly higher CAIs (Figure S5D, E).

For the analyses in Figures 3C, D and S5B, C we calculated the CAI (to human codon usage) for each branch, according to the maximum-likelihood ancestral sequence reconstruction (PAML v4.7) and converted it to a Z-score (by subtracting the mean CAI across branches and dividing by the standard deviation). Z-scores can be less than one, so the minimum Z-score in each tree was subtracted from the Z-score for each branch of the tree, such that the displayed branch have length zero or greater. These analyses were performed using the Batch 1 dataset, and both trees were rooted on Pinneo (not shown).

A Kolmogorov-Smirnov test comparing Nigerian and non-Nigerian branches shows a significant difference in CAI ($D = 0.48$, P-value = 0.0005), but the non-independence among branches violates the assumptions of the test. For this reason we used empirical P-values, based on permutations accounting for the tree structure. The significance of the difference in CAI between Nigerian and non-Nigerian sequences was tested by permuting the reconstructed sequence changes across the maximum-likelihood tree topology and testing whether the permuted sequences had significantly different distributions of CAI between Nigerian and non-Nigerian branches (Kolmogorov-Smirnov test, P-value < 0.05). The permutation was repeated 1000 times to obtain an empirical P-value equal to the fraction of permutations showing a significant difference between Nigerian and non-Nigerian CAI.

Intra-host variant calling. For each sequenced sample, reads were realigned to the consensus sequence using Novoalign as described above and iSNVs were called using mpileup with the following parameters: samtools mpileup -Q 0 -B -q 1 -d 10000 and VarScan v2.3 (Koboldt et al., 2012) with the following parameters: varscan.jar pileup2snp --min-reads2 5 --min-var-freq 0.01 --p-value 0.1 --min-coverage 5 --min-avg-qual 5. Stringent post-call filtering were applied to remove false calls using the following variables:

- Minimum overall coverage of 5x.
- Minimum depth for variant calls at each position of 5.
- Minimum iSNV frequency of 5%
- Maximum strand-bias difference of 10-fold between variant and reference alleles
- Minimum base-quality call of Q25
- Maximum base-quality difference between variant and consensus call of Q5.

For comparison analyses (Figure S5J) GATK v2.1 was used as well with the following parameters: GATK calling: -baq OFF --useOriginalQualities -dt NONE --

min_base_quality_score 20 -ploidy 10 -stand_call_conf 80.0 -stand_emit_conf 30.0 -A AlleleBalance; GATK filtering: -l ERROR --filterExpression "(MQ0/DP) > 0.20 || BaseQRankSum < -10.0 || QD < 1.0 || MQRankSum < -4.0 || ReadPosRankSum < -4.0" --filterName LowConfidence --filterExpression "DP < 20 || (DP*(1-ABHet)) < 5" --filterName LowCoverage --filterExpression "ABHet > 0.95" --filterName LowFrequency --filterExpression "SB < -1.0e+09 || SB > 1.0e+09 || FS > 5" --filterName StrandBias.

We experimented with different minMAFs and found that the 5% cutoff used in this study gave reproducible results across platforms and replicates (Figures S5J and S6). The filtered iSNV calls (5% minMAF) can be found in File S1. In contrast, we found that calling iSNVs at 1% or below – even given very high coverage – required multiple technical replicates and independent library preparations (Gire et al., 2014). This is likely due to the random incorporation of RT-PCR and PCR errors during library construction.

Validation of intra-host variants using 454 and Sanger sequencing. Two samples with high numbers of iSNVs were selected for variant validation using 454 sequencing (rodent Z0947) or Sanger sequencing (patient G733). For 454 sequencing the entire LASV genome was amplified using four sets of 'DemiHemi' primers (Table S1, primers tab, L1, L2, L3a, Sa) and sequenced on the 454 platform using previously described protocols (Lennon et al., 2010; Henn et al., 2012). For Sanger sequencing a panel of 17 M13-tagged primer sets (Table S1, primers tab) was created and amplicons were generated and sequenced using standard Sanger sequencing protocols (GeneWiz).

Comparison of iSNV rates. In order to compare levels of polymorphism between samples sequenced with varying coverage, we computed an iSNV rate for each sample: the number of iSNVs called at sites with sufficient coverage (N) divided by the total number of sites with sufficient coverage (N). The probability of calling an iSNV (power) was modeled with a binomial distribution to calculate the minimum sufficient coverage (N) at a site to call an iSNV with a given minor allele frequency (MAF), with the minor allele present on at least 5 reads. At 90% iSNV calling power, N ≥ 926 is sufficient for MAF ≥ 1%, N ≥ 184 for MAF ≥ 5%, N ≥ 91 for MAF ≥ 10%, N ≥ 70 for MAF ≥ 12.5%. Specifically, in Figure 5C we calculated the iSNV rate using only sites with sufficient coverage to call iSNVs with 80% calling power in each minor allele frequency (MAF) bin and tested the difference between human and *M. natalensis* separately within each MAF bin.

Rarefaction analysis. All LASV reads were extracted from aligned BAM files with duplicates removed and multiple runs from the same sample were combined. Depending on the average coverage the various samples were then downsampled to an average of 500x, 200x, 100x or 50x for 20 replicates. Each replicate was then subsampled in fractions of 0.1 (0.1 to 1.0) and the subsampled reads were aligned to their own LASV consensus sequences using Novoalign v2.08.02. Intra-host variants were then called with VarScan v2.3 as described above.

Assessment of transition/transversion ratios. Higher transition/transversion ratio does not explain the high CAI in Sierra Leone. The Sierra Leonean LASV population is relatively young (Figure 2G), and has a significantly different mutation spectrum, enriched in transitions relative to Nigerian strains (Figure S5L). However, changes in synonymous codon preference between Sierra Leone and Nigeria are not enriched in transition mutations, as might be predicted if the change in CAI had been driven by mutational biases. For each four-fold degenerate amino acid, we computed an odds ratio statistic to assess the relative preference for each synonymous codon between Sierra Leone and Nigeria. We ranked the absolute value of the odds ratios for each of the 5 four-fold degenerate amino acids. If synonymous changes involving transitions were consistently highly ranked, this would suggest that they are responsible for the differences in codon preference between Sierra Leone and Nigeria. However, we observed that transitions (A/G or C/T) had the top-ranked odds ratio for only 1 out of 5 amino acids (expected (2/6) * 5 = 1.67) on the S segment and 3 out of 5 on the L segment. This suggests that a transition bias is not entirely responsible for the difference in codon preference between Sierra Leone and Nigeria. We therefore cannot exclude the role of selection in shaping CAI; however more work will be needed to determine whether high or low CAI is more adaptive for LASV.

Intra-host selection analysis. To investigate signatures of natural selection within infected hosts, we examined the patterns of LASV iSNV substitutions observed in patients and *M. natalensis* using the McDonald-Kreitman (MK) framework. This test has classically been used to test for selection at the protein level by calculating the ratio of the number of nonsynonymous and synonymous substitutions (N/S) within species versus between species (McDonald and Kreitman, 1991). More recently, it has been extended to distinguish selective pressures within versus between populations of viruses (Renzette et al., 2011; Bhatt et al., 2010). Using this framework, we compared variation in LASV within and between hosts, using a measure called the Neutrality Index (NI) (Rand and Kann, 1996), defined as the N/S ratio between iSNVs divided by N/S between consensus sequences found in different hosts. Under neutral evolution, these ratios should be equal, yielding NI = 1. A significantly high NI is generally interpreted as evidence for purifying (negative) selection between hosts; however it can also indicate diversifying (positive) selection within hosts. NI is the reciprocal of the more commonly used Fixation Index (McDonald and Kreitman, 1991). Because most of our discussion focuses on evolution within hosts, NI is the more intuitive statistic to use here.

The MK framework has been previously applied to quantify selective pressures in viruses (Bhatt et al., 2010), but care must be taken to accurately estimate between-host N and S in rapidly evolving viral populations, where multiple substitutions may occur at the same site. Therefore, we only considered substitutions on leaf branches, which generally contain only a few substitutions, unlike long internal branches, which are often saturated with multiple synonymous substitutions. Between-host variants were inferred as changes between a consensus sequence (leaf branch) and all other sequences in the same phylogenetic cluster. Phylogenetic clusters were defined as monophyletic groups of the maximum likelihood phylogeny with a maximum of 0.01 substitutions/site from the root of the group to the tips. Only branches containing iSNVs were included in the between-host counts, although including all leaf branches did not significantly change the between-host N/S estimate (data not shown). The significance of the deviation of NI from the neutral expectation was evaluated by permuting the 2x2 contingency tables across genes and samples, keeping the row and column marginals constant, as previously described (Shapiro et al., 2007). The *P*-value of the observed NI was calculated as the fraction of 10,000 permutations that yielded an NI greater or equal to the observed NI (Figure 5D and E). The expected NI was defined as the mean permuted NI, and the normalized NI defined as observed divided by expected NI. In Figures 5, 6 and 7 we used a minMAF of 5%. In Figure 5, the N and S counts are normalized by the number of nonsynonymous or synonymous sites to obtain dN and dS, respectively, and calculate the dN/dS ratio.

Epitope prediction. B cell epitopes were predicted in protein-coding regions of LASV consensus sequence using BCPREDS with 75% specificity and an epitope length of 20 amino acids (El-Manzalawy et al., 2008b). We also predicted B cell epitopes in LASV GPC using FBCPred (EL-Manzalawy et al., 2008a), which allows for flexible length epitope prediction and obtained very similar results (Figure S7I). Comparing the predicted B cell epitopes to a smaller subset of experimentally determined epitopes (GPC only) available in the ViPR database (as of April 6th, 2015, http://www.viprbrc.org/), we found that five out of eight of the experimental B cell epitopes had overlap with the predicted epitopes. Epitopes were predicted separately on each consensus sequence, and overlaid with the iSNVs from the matched sample. The overlap between iSNVs and epitopes in each protein was assessed using a binomial test, with binomial probability p equal to the fraction of amino acids in the protein covered by predicted epitopes. The epitope predictions were highly consistent across different consensus sequences, therefore a single average p was used for each protein. The iSNVs falling within or outside predicted epitopes were summed across consensus sequences. This assumes independence of iSNVs, which is a reasonable assumption given that there is little evidence for strong linkage between adjacent iSNVs (Table S2). Moreover, the vast majority of samples contain only 1 iSNV; therefore any linkage effects are expected to be minimal.

T cell epitope predictions were performed using NetCTL (Larsen et al., 2007) as implemented on the ViPR website (http://www.viprbrc.org/). A cutoff score of 1.5 was used with Josiah as the reference and epitopes were calculated using all available MHC class I supertypes.

The predicted B cell and T cell epitopes can be found in File S1.

iSNV fixation analysis. Fixation of iSNVs was considered separately, first, across all sequenced samples, and second, between a pair of rodents captured on the same day from the same household, for which recent transmission was suspected (Z0947, Z0948). Their consensus sequences were nearest-neighbors on the inferred LASV phylogeny (Files S1-3), suggesting a recent transmission event. We further suspected that transmission occurred from Z0947 to Z0948 based on two observations. First, Z0947 contained a substantially more diverse LASV population (27 iSNVs at 10% MAF) compared to Z0948 (3 iSNVs at 10% MAF). This suggests a longer infection period in Z0947, and a more recent transmission to Z0948. Second, all 3 iSNVs in Z0948 are nonsynonymous, resulting in a high N/S ratio (compared to an N/S ratio of 8/19 =

0.42 in Z0947). This suggests that slightly deleterious nonsynonymous variants are circulating in Z0948, consistent with a recent population bottleneck (expected during transmission) and/or a relatively recently established viral population in Z0948. While these data support a "Z0947 to Z0948" transmission scenario, we cannot formally exclude other scenarios (Figure S7H). To ensure that iSNVs in Z0947 were not also present at low frequencies in Z0948 (and *vice versa*), we manually inspected the aligned sequence reads. Therefore, either the major or minor allele at each Z0947 iSNV must necessarily be fixed (within the resolution of our sequencing depth) in the Z0948 consensus sequence.

Across all sequenced samples, an iSNV was defined to be 'fixed' if the minor allelic variant was observed in one or more consensus sequences in the entire dataset. The same analysis was performed, focusing on iSNVs in Z0947, either with Z0948 as the only outgroup, or with Z0948 plus two additional, more distant outgroups (G1190 and G1727), in order to account for different possible transmission scenarios (Figure S7H). Specifically, we considered only polymorphic sites (iSNVs or fixed differences between consensus sequences) for which minor/major Z0947 iSNVs could be unambiguously assigned as derived/ancestral with reference to additional outgroups (G1190 and G1727). This resulted in the exclusion of 3 ambiguous sites.

The assumption that iSNVs can be counted independently is justified based on the general lack of linkage between adjacent iSNVs on the same sequenced read (Table S2). Nevertheless, as a precaution to guard against undetected linkage effects in the data, we repeated the analysis of minor iSNV alleles, only counting a maximum of one nonsynonymous (N) or synonymous (S) iSNV (chosen at random) per sample per segment, from Batch 1 sequences. In this way, each potentially linked haplotype is only counted once, yielding the following counts: N unfixed = 10, N fixed = 18, S unfixed = 0, S fixed = 40 (Fisher's exact test: Odds ratio > 22, *P*-value = 4.5E-5).

Linkage between iSNVs. In a few cases, nearby iSNV sites were sequenced on the same sequencing read (or read pair), allowing us to assess linkage between them. We gathered all such iSNV pairs, requiring that each iSNV have a quality score ≥ 20 and minMAF of 10%. We then counted major-major, major-minor, minor-major and minor-minor allele combinations (haplotypes) and computed the D_A' linkage measure, which ranges from 0 (unlinked) to 1 (perfectly linked), taking allele frequencies into account (Hedrick, 1987; Kalinowski and

Hedrick, 2001). On average, linkage was relatively weak (mean $D_A' \approx 0.4$; Table S2) but the number of reads spanning two iSNVs was generally also quite small; therefore we hesitate to draw firm conclusions about the extent of linkage in our dataset.

Luciferase constructs containing GBlocks. We designed 750 bp GBlocks (Integrated DNA Technologies) containing 699 bp of the 5’ end of the LASV NP gene, each for 10 Nigerian strains and 10 Sierra Leonean strains, along with 21-25 bps of flanking sequence at either end homologous to the m6pgfp vector (Andersen et al., 2008). As positive controls, GBlocks were ordered with human codon-optimized versions of LASV NP based on ISTH0073 (Nigerian) and G1442 (Sierra Leonean). Sequences were codon-optimized using the IDT Codon Optimization Tool (Integrated DNA Technologies) and the optimized sequences were between 70-75% identical to the parent sequences at the nucleotide level (100% at the amino acid level). GBlocks were assembled into PCR linearized m6pgfp vector using Gibson Assembly (New England Biolabs) to create an NP-firefly Luciferase (fLuc) fusion driven by the MMLV LTR. A single sequence-verified assembled clone was propagated overnight and plasmid DNA extracted using the Plasmid Plus Midi kit (Qiagen). These plasmids were designated ‘pKGAC-fLuc.’

To generate pKGAC-gLuc series constructs, each pKGAC-fLuc construct was double digested with EcoRI and NotI (New England Biolabs), heat inactivated, and treated with Antarctic phosphatase (New England Biolabs). Vector backbones were subsequently gel purified with the QIAquick Gel Extraction Kit (Qiagen). A DNA insert containing Gaussia-Dura luciferase (henceforth referred to as 'gLuc') was generated by PCR (95° C 2 minutes; 95° C 20 seconds, 55° C 15 seconds, 70° C 12 seconds, 35 cycles; 70° C 2 minutes) of pTK-Gaussia-Dura Luc (Thermo Scientific) with the primers gLuc_FW and gLuc_RV (Table S1, primers tab), which contain regions of homology to the template as well as 5' overhangs generating EcoRI and NotI restriction sites. The PCR product was run on an agarose gel and purified with the QIAquick Gel Extraction Kit. Subsequently, the product was double digested with EcoRI and NotI to expose sticky ends, and purified with QIAquick Spin Columns (Qiagen). The gLuc insert was ligated in-frame to each vector backbone with T4 DNA ligase (New England Biolabs), and subsequent ligation products were used to transform chemically competent NEB 10β *E. coli* (New England Biolabs). Transformants were plated, and individual colonies were picked and sequenced to verify correct, in-frame insertion of the gLuc insert. These plasmids were designated ‘pKGAC-gLuc.’

fLuc Transfection, luminescence, and RT-qPCR. 100 µl of HEK293 cells (Life Technologies) were seeded into three opaque white (for detecting luminescence) 96-well plates, grown to 60-80% confluency, and transfected with Lipofectamine LTX & PLUS Reagent using the manufacturer's guidelines (Life Technologies). Briefly, each pKGAC-fLuc plasmid was incubated for 5 minutes at room temperature with pGL4.74-hRluc (Promega), which encodes the Renilla luciferase (rLuc) as an internal control, and PLUS Reagent. Next, Lipofectamine LTX Reagent was added, and the mixture was incubated for another 30 minutes at room temperature. The mixture was aliquoted into cells in the three 96-well plates. Each replicate contained 101.5 ng pKGAC-fLuc DNA, 10 ng pGL4.74-hRluc, 0.1 µl PLUS Reagent, and 0.4 µl Lipofectamine LTX Reagent, administered to the cell media in a 20 µl mixture.

To detect NP-fLuc fusion protein expression, a luciferase assay was performed using the Dual-Glo Luciferase Assay System (Promega). 16-20 hours after transfection, 50 µl of media was removed from each well of the 3 opaque white 96-well plates. 50 µl Dual-Glo Luciferase Reagent with substrate was added directly to the culture media of each well, mixed vigorously to ensure full cell lysis, and incubated for 10 minutes at room temperature. Firefly luminescence was detected on a Spectramax L with 5 seconds integration and 470 nm calibration wave. Next, 50 µl Dual-Glo Stop & Glo Reagent with substrate (Promega) was added to quench firefly luminescence and initiate renilla luminescence. Following 10 minutes incubation at room temperature, renilla luminescence was detected on a Spectramax L with 5 seconds integration and 470 nm calibration wave. In data analysis, firefly luminescence was normalized against renilla luminescence (fLuc/rLuc) for each sample. The three technical replicate plates were then averaged.

The entire experiment was performed in biological triplicate to produce the final data. For each of the 3 biological replicates, the fLuc/rLuc of each sample was then normalized against the average fLuc/rLuc of all non-control samples on that plate.

***In vitro* transcription.** pKGAC-gLuc series plasmids were used to generate $NP_{1\text{-}699}$-gLuc RNA for *in vitro* translation assays. First, DNA amplicons containing the T7 RNA polymerase promoter (T7Rpo) sequence upstream of NP_{1-699}-gLuc were generated by PCR (94° C 2 minutes; 94° C 15 seconds, 40° C 15 seconds, 72° C 25 seconds, 5 cycles; 94° C 15 seconds, 55° C 15 seconds, 72° C 25 seconds, 30 cycles; 70° C 1 minute) of pKGAC-gLuc plasmids with the primers pKGAC-gLuc_FW and pKGAC-gLuc_RV (Table S1, primers tab), which contain regions of homology to the template as well as a 5' overhang to generate T7Rpo in the sense direction. PCR products were gel purified with the QIAguick Gel Extraction Kit (Qiagen).

In addition, to investigate the translation of LASV GPC, we designed 1,325 bp GBlocks (Integrated DNA Technologies) containing the T7Rpo sequence, 736 bp of the 5' end of the LASV GPC gene, each for 5 Nigerian strains and 5 Sierra Leonean strains, and the sequence encoding gLuc. DNA amplicons were generated by PCR (94° C 2 minutes; 94° C 15 seconds, 58° C 15 seconds, 72° C 30 seconds, 40 cycles; 72° C 1 minute) with the primers pKGAC-gLuc_FW and pKGAC-gLuc_RV (Table S1, primers tab), and gel purified.

Next, DNA amplicons were used as templates for *in vitro* transcription by T7 RNA polymerase using a reduced version of the MEGAscript T7 Transcription Kit (Life Technologies) per the manufacturer's protocol. Briefly, each 20 µl reaction contained ~250-400 ng DNA template, 1X reaction buffer, 50 nmol of each NTP, and 2/3 µl of enzyme mix, and transcription proceeded at 37° C for 4 hours. To remove template DNA, samples were treated with 2 U TURBO DNase (Life Technologies) at 37° C for 20-30 minutes. For each of the 20 NP and 10 GPC DNA amplicons, *in vitro* transcription was performed in biological triplicate, resulting in 60 NP and 30 GPC RNA samples, in addition to codon optimized/de-optimized controls. Each *in vitro* transcription product was purified with Agencourt RNAClean XP beads (Beckman Coulter).

***In vitro* translation and gLuc detection.** Purified *in vitro* transcription products were used as templates for *in vitro* translation using a reduced version of the Retic Lysate IVT Kit (Life Technologies). Briefly, we set up a modified reaction per the manufacturer's protocol except using 4.25 µl of reticulocyte lysate per 25 µl reaction (25% of the recommended lysate amount). 2.5 µg of each *in vitro* transcribed RNA was subjected to *in vitro* translation at 30° C for 21 hours. Translation was also performed in biological triplicate. Immediately following translation, each sample was incubated with 25 µl of coelenterazine substrate from the Pierce Gaussia Luciferase Glow Assay Kit (Thermo Scientific) at room temperature for 10 minutes.

Luminescence was detected on a Spectramax L with 0.5 seconds integration and 395 nm calibration wave.

LASV GPC cloning and expression. LASV Josiah GPC was used as a control. Several LASV GPC iSNV mutants based on the LM395 and LM776 *M. natalensis* samples were created as GBlocks (Integrated DNA Technologies) and cloned into the pcDNA3.1+zeo_intA vector for high level expression in mammalian cells (Life Technologies). The following mutants were made: LM395^{WT}, LM395^{D89N}, LM395^{L113I}, LM395^{N114D}, LM395^{R161M}, LM395^{R248K}, LM776^{WT}, and LM776^{E67Q}. All iSNV mutants were located in the GP1 part of LASV GPC and all LM395 mutants overlapped with predicted or experimental B cell epitopes. The cloning strategy was identical to that used for generation of the LASV Josiah GPC construct and clones were verified by DNA Sanger sequencing.

Endotoxin-free plasmid DNAs were generated from transformed E. coli TOP10 clones, quantitated and used in transient transfection studies. HEK293 cells were seeded the day before transfection in Poly-D-Lysine treated 6-well polystyrene plates in DMEM, 10% qFBS, 2mM L-Gln. On the day of transfection, fresh medium was exchanged and cells were incubated with transfection mix contain pre-optimized ratios of DNA and PEI (PolyPlus), according to manufacturer's recommendations. Twenty-four hours after transfection cells were gently washed twice with PBS and incubated with fresh PBS for 15 minutes at room temperature to allow for gentle dislodging without mechanical disruption or enzymatic detachment. Cells were centrifuged for 5 minutes at 250x*g*, washed twice in FACS buffer (1X PBS, pH7.4, 2% FBS, 0.05% NaN3), and used in binding experiments. Cells (10^5/assay) were incubated in polypropylene 96-well U-bottom plates with prediluted LASV human monoclonal or control antibodies in FACS buffer, for 20 minutes on ice. Cells were washed twice by centrifugation in FACS buffer and incubated with diluted Goat anti-human IgG(H+L)-Alexa488 secondary reagent in FACS buffer for an additional 20 minutes on ice. Following 2 additional centrifugation washes, cells were resuspended in PI buffer (FACS buffer with 1 μg/mL propidium iodide). Ten thousand live cell events (PI negative) per samples were collected and analyzed in a BD Accuri C6 cytometer, and mean channel fluorescence values were derived.

To verify expression of LASV GPC constructs in HEK293 cells, extent of GP1 and GP2 cleavage, and sGP1 secretion western blots were performed on cell extracts and supernatants.

Briefly, cell extracts were prepared with the Sigma Mammalian Cell Lysis Kit (Sigma), and an equivalent input of protein from each extract was resolved on Bis-Tris 4-20% SDS-PAGE gels (Life Technologies), transferred to nitrocellulose membranes, and probed with LASV glycoprotein-specific murine antibodies raised against irradiated LASV antigen. Blots were probed with a Goat F(ab')2 anti-mouse IgG(H+L)-HRP secondary reagent and LumiGlo chemiluminescent substrate. All images were captured on a GE Image Quant LAS4000 gle docking station. Similarly, equal volumes of centrifugation cleared supernatants from each transfection were resolved by SDS-PAGE and probed in western blots with the same reagents.

Human LASV-specific monoclonal antibodies. The human monoclonal antibodies (12.1F, 19.7E, 25.6A, 36.1F, 37.7H, and 37.2D) used in these studies were derived from PBMC of LF convalescent patients. Briefly, PBMC were isolated from buffy coats obtained from Ficoll-Isopaque centrifugation gradients, prepared at KGH. Cells were cryopreserved by slow cooling to -80^{o} C, and shipped to Tulane University in IATA-approved dry shippers. PBMC were subsequently thawed, enriched for Pan B cells and seeded at nearly clonal densities in flat bottom 96 well plates, under activation conditions that specifically induced B cell proliferation. All emerging clones were screened for the presence of human antibodies specific for the LASV Josiah GPC, by ELISA and direct neutralization assays using a LASV GPC pseudotyped lentiviral particle system.

Antibodies with demonstrated binding and/or neutralization were cloned from corresponding cells by PCR amplification of light and heavy chain cDNAs using universal human IgG oligonucleotides and high fidelity DNA polymerases. The heavy and light chains were cloned in expression vectors, verified by DNA sequencing, and expressed in transiently transfected HEK293 cells. Recombinant antibodies that exhibited binding and neutralization to LASV GPC expressed on the surface of HEK293 cells were subsequently cloned in dual expression mammalian vectors for generation of stable NS0 (null secreting) murine myeloma cell lines (CHOLCelect). These studies were performed with NS0 cell produced antibodies. All antibodies were purified from chemically defined serum free stable NS0 cultures by Protein A chromatography, dialyzed in a proprietary buffer for injection, and diluted in FACS buffer for cytometry assays.

Limited mapping studies have been performed and it is believed that 19.7E binds an epitope in LASV GP1 around position 114 and 12.1F is in the same complementation group as 19.7E, but off to one side where it binds a different epitope. All other antibodies are believed to bind epitopes that lie within GP2 but may cover quaternary epitopes that also require sites in GP1.

SUPPLEMENTAL REFERENCES

Adiconis, X., Borges-Rivera, D., Satija, R., DeLuca, D. S., Busby, M. A., Berlin, A. M., Sivachenko, A., Thompson, D. A., Wysoker, A., Fennell, T., Gnirke, A., Pochet, N., Regev, A., and Levin, J. Z. (2013). Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat Methods *10*, 623-629.

Agarwala, R., and Rotmistrovsky, K. BMTagger: Best Match Tagger for removing human reads from metagenomics datasets.

Andersen, K. G., Butcher, T., and Betz, A. G. (2008). Specific immunosuppression with inducible Foxp3-transduced polyclonal T cells. PLoS Biol *6*, e276.

Bhatt, S., Katzourakis, A., and Pybus, O. G. (2010). Detecting natural selection in RNA virus populations using sequence summary statistics. Infect Genet Evol *10*, 421-430.

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyras, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H. R., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, K. C., Cameron, G., Durbin, R., Cox, A., Hubbard, T., and Clamp, M. (2004). An overview of Ensembl. Genome Res *14*, 925-928.

Bowen, M. D., Rollin, P. E., Ksiazek, T. G., Hustad, H. L., Bausch, D. G., Demby, A. H., Bajani, M. D., Peters, C. J., and Nichol, S. T. (2000). Genetic diversity among Lassa virus strains. J Virol *74*, 6992-7004.

Branco, L. M., Boisen, M. L., Andersen, K. G., Grove, J. N., Moses, L. M., Muncy, I. J., Henderson, L. A., Schieffellin, J. S., Robinson, J. E., Bangura, J. J., Grant, D. S., Raabe, V. N., Fonnie, M., Zaitsev, E. M., Sabeti, P. C., and Garry, R. F. (2011a). Lassa hemorrhagic fever in a late term pregnancy from northern Sierra Leone with a positive maternal outcome: case report. Virol J *8*, 404.

Branco, L. M., Grove, J. N., Boisen, M. L., Shaffer, J. G., Goba, A., Fullah, M., Momoh, M., Grant, D. S., and Garry, R. F. (2011b). Emerging trends in Lassa fever: redefining the role of immunoglobulin M and inflammation in diagnosing acute infection. Virol J *8*, 478.

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972-1973.

CBRC (2012). Lastal available online.

Drummond, A. J., Ashton, B., Cheung, M., Heled, J., Kearse, M., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T., and Wilson, A. (2010). Geneious v5.0, Available from http://www.geneious.com.

Drummond, A. J., Ho, S. Y., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. PLoS Biol *4*, e88.

Drummond, A. J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol *7*, 214.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol *22*, 1185-1192.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460-2461.

EL-Manzalawy, Y., Dobbs, D., and Honavar, V. (2008a). Predicting flexible length linear B-cell epitopes. 7th International Conference on Computational Systems Bioinformatics 121-131.

El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2008b). Predicting linear B-cell epitopes using string kernels. J Mol Recognit *21*, 243-255.

Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S., Park, D. J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., Wohl, S., Moses, L. M., Yozwiak, N. L., Winnicki, S., Matranga, C. B., Malboeuf, C. M., Qu, J., Gladden, A. D., Schaffner, S. F., Yang, X., Jiang, P. P., Nekoui, M., Colubri, A., Coomber, M. R., Fonnie, M., Moigboi, A., Gbakie, M., Kamara, F. K., Tucker, V., Konuwa, E., Saffa, S., Sellu, J., Jalloh, A. A., Kovoma, A., Koninga, J., Mustapha, I., Kargbo, K., Foday, M., Yillah, M., Kanneh, F., Robert, W., Massally, J. L., Chapman, S. B., Bochicchio, J., Murphy, C., Nusbaum, C., Young, S., Birren, B. W., Grant, D. S., Scheiffelin, J. S., Lander, E. S., Happi, C., Gevao, S. M., Gnirke, A., Rambaut, A., Garry, R. F., Khan, S. H., and Sabeti, P. C. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science *345*, 1369-1372.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol *29*, 644-652.

Hedrick, P. W. (1987). Estimation of the rate of partial inbreeding. Heredity (Edinb) *58*, 161-166.

Henn, M. R., Boutwell, C. L., Charlebois, P., Lennon, N. J., Power, K. A., Macalalad, A. R., Berlin, A. M., Malboeuf, C. M., Ryan, E. M., Gnerre, S., Zody, M. C., Erlich, R. L., Green, L. M., Berical, A., Wang, Y., Casali, M., Streeck, H., Bloom, A. K., Dudek, T., Tully, D., Newman, R., Axten, K. L., Gladden, A. D., Battis, L., Kemper, M., Zeng, Q., Shea, T. P., Gujja, S., Zedlack, C., Gasser, O., Brander, C., Hess, C., Gunthard, H. F., Brumme, Z. L., Brumme, C. J., Bazner, S., Rychert, J., Tinsley, J. P., Mayer, K. H., Rosenberg, E., Pereyra, F., Levin, J. Z., Young, S. K., Jessen, H., Altfeld, M., Birren, B. W., Walker, B. D., and Allen, T. M. (2012). Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog *8*, e1002529.

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. Genome Res *21*, 1543-1551.

Kalinowski, S. T., and Hedrick, P. W. (2001). Estimation of linkage disequilibrium for loci with multiple alleles: basic approach and an application using data from bighorn sheep. Heredity (Edinb) *87*, 698-708.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res *30*, 3059-3066.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res *22*, 568-576.

Kosakovsky Pond, S. L., Murrell, B., Fourment, M., Frost, S. D., Delport, W., and Scheffler, K. (2011). A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol *28*, 3033-3043.

Larsen, M. V., Lundegaard, C., Lamberth, K., Buus, S., Lund, O., and Nielsen, M. (2007). Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. BMC Bioinformatics *8*, 424.

Lennon, N. J., Lintner, R. E., Anderson, S., Alvarez, P., Barry, A., Brockman, W., Daza, R., Erlich, R. L., Giannoukos, G., Green, L., Hollinger, A., Hoover, C. A., Jaffe, D. B., Juhn, F., McCarthy, D., Perrin, D., Ponchner, K., Powers, T. L., Rizzolo, K., Robbins, D., Ryan, E., Russ, C., Sparrow, T., Stalker, J., Steelman, S., Weiand, M., Zimmer, A., Henn, M. R., Nusbaum, C., and Nicol, R. (2010). A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. Genome Biol *11*, R15.

Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods *7*, 709-715.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, W. L., and Drummond, A. J. (2012). Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. Mol Biol Evol *29*, 751-761.

Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M., and Usadel, B. (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res *40*, W622-W627.

Loytynoja, A., and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A *102*, 10557-10562.

Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D., and Lefeuvre, P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics *26*, 2462-2463.

Matranga, C. B., Andersen, K. G., Winnicki, S., Busby, M., Gladden, A. D., Tewhey, R., Stremlau, M., Berlin, A., Gire, S. K., England, E., Moses, L. M., Mikkelsen, T. S., Odia, I., Ehiane, P. E., Folarin, O., Goba, A., Kahn, S., Grant, D. S., Honko, A., Hensley, L., Happi, C., Garry, R. F., Malboeuf, C. M., Birren, B. W., Gnirke, A., Levin, J. Z., and Sabeti, P. C. (2014). Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. Genome Biol *15*, 519.

McCormick, J. B., King, I. J., Webb, P. A., Scribner, C. L., Craven, R. B., Johnson, K. M., Elliott, L. H., and Belmont-Williams, R. (1986). Lassa fever. Effective therapy with ribavirin. N Engl J Med *314*, 20-26.

McDonald, J. H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature *351*, 652-654.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297-1303.

Nagarajan, N., and Kingsford, C. (2011). GiRaF: robust, computational identification of influenza reassortments via graph mining. Nucleic Acids Res *39*, e34.

Novocraft (2012). Novoalign available online.

Olschlager, S., Lelke, M., Emmerich, P., Panning, M., Drosten, C., Hass, M., Asogun, D., Ehichioya, D., Omilabu, S., and Gunther, S. (2010). Improved detection of Lassa virus by reverse transcription-PCR targeting the 5' region of S RNA. J Clin Microbiol *48*, 2009-2013.

Pond, S. L., Frost, S. D., and Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. Bioinformatics *21*, 676-679.

Puigbo, P., Bravo, I. G., and Garcia-Vallve, S. (2008a). CAIcal: a combined set of tools to assess codon usage adaptation. Biol Direct *3*, 38.

Puigbo, P., Bravo, I. G., and Garcia-Vallve, S. (2008b). E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). BMC Bioinformatics *9*, 65.

Rand, D. M., and Kann, L. M. (1996). Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Mol Biol Evol *13*, 735-748.

Renzette, N., Bhattacharjee, B., Jensen, J. D., Gibson, L., and Kowalik, T. F. (2011). Extensive genome-wide variability of human cytomegalovirus in congenitally infected infants. PLoS Pathog *7*, e1001344.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet *16*, 276-277.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863-864.

Shapiro, J. A., Huang, W., Zhang, C., Hubisz, M. J., Lu, J., Turissini, D. A., Fang, S., Wang, H. Y., Hudson, R. R., Nielsen, R., Chen, Z., and Wu, C. I. (2007). Adaptive genic evolution in the Drosophila genomes. Proc Natl Acad Sci U S A *104*, 2271-2276.

Shimodaira, H., and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol *16*, 1114-1116.

Stamatakis, A., Ludwig, T., and Meier, H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics *21*, 456-463.

Stremlau, M. H., Andersen, K. G., Folarin, O. A., Grove, J. N., Odia, I., Ehiane, P. E., Omoniwa, O., Omoregie, O., Jiang, P. P., Yozwiak, N. L., Matranga, C. B., Yang, X., Gire, S. K., Winnicki, S., Tariyal, R., Schaffner, S. F., Okokhere, P. O., Okogbenin, S., Akpede, G. O., Asogun, D. A., Agbonlahor, D. E., Walker, P. J., Tesh, R. B., Levin, J. Z., Garry, R. F., Sabeti, P. C., and Happi, C. T. (2015). Discovery of novel rhabdoviruses in the blood of healthy individuals from west Africa. PLoS Negl Trop Dis *9*, e0003631.

Suchard, M. A., Weiss, R. E., and Sinsheimer, J. S. (2003). Testing a molecular clock without an outgroup: derivations of induced priors on branch-length restrictions in a Bayesian framework. Syst Biol *52*, 48-54.

Institute, T. B. (2012a). FastQC avaliable online.

Institute, T. B. (2012b). Picard available online.

Wertheim, J. O., Chu, D. K., Peiris, J. S., Kosakovsky Pond, S. L., and Poon, L. L. (2013). A case for the ancient origin of coronaviruses. J Virol *87*, 7039-7045.

Wertheim, J. O., and Kosakovsky Pond, S. L. (2011). Purifying selection can obscure the ancient age of viral lineages. Mol Biol Evol *28*, 3355-3365.